Improving sentence embedding with sentence relationships from word analogies

ZHANG Qixuan

Graduate School of Information, Production, and System, Waseda University

July 17, 2023

早稲田大学 情報生産システム研究科
Graduate School of Information, Production and Systems, Waseda University

Outline

Word/Sentence Embedding

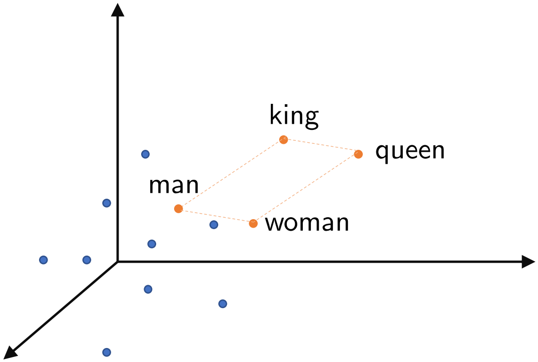
Represents words or sentences as vectors. These representations are used in:

- document retrieval
- sentiment analysis
- machine translation
- ......

Key point: Representing the meaning of the text

Word/Sentence Embedding

Word Embedding Space



Word/Sentence Embedding

Sentence Embedding Space

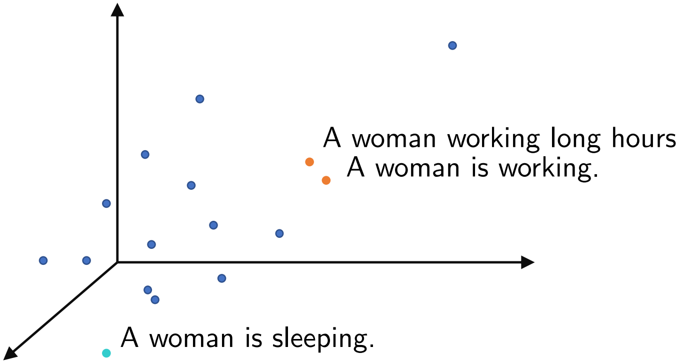


Figure: Visualized sentence embedding space

Sentence embedding methods

Sentence embedding learned from context

- Skip-thoughts (Kiros et al., 2015)
- Quick-thoughts (Logeswaran and Lee, 2018)

Sentence embedding methods

Sentence embedding learned from relations between sentences

- InferSent (Conneau et al., 2017)
- Sentence-BERT (Reimers and Gurevych, 2019)
- SimCSE (Gao et al., 2021)

Downstream Evaluation

Table: Evaluation results of sentence embeddings. Table copied from (Li et al., 2022). Methods based on sentence relationships perform better.

	STS12-16	MR	CR	MPQA	SST2
Skip-thoughts	43.00	76.56	79.88	86.91	82.16
Quick-thoughts	51.00	80.33	83.52	89.32	85.23
SBERT-large-NLI	**75.00**	84.81	90.92	90.23	90.85
SRoBERTa-large-NLI	74.00	**87.07**	**91.41**	**90.60**	**92.25**

Natural Language Inference (NLI) Corpus

Table: Example extracted from the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015)

Premise	Hypotheses	Label
A woman working long hours.	A woman is working.	entailment
	A woman is working in a factory.	neutral
	A woman is sleeping.	contradiction

Natural Language Inference (NLI) Corpus

Relation between sentences $\rightarrow$ World knowledge

Natural Language Inference (NLI) Corpus

Construction of the SNLI corpus (Bowman et al., 2015):

- Crowdsourcing usinh Amazon Mechanical Turk
- About 2,500 human workers
- Premise: Flickr30k (also crowdsourcing work)
- Workers wrote hypothesis sentences for premise

Our main work

- Generation of sentence relationship data: DSBATS-sn (Definition Sentences from BATS with semantic network)
- Evaluation of the generated sentence relationships, verification of validity of DSBATS-sn

Contribution

- A new method to obtain the relationship between sentences automatically with more diverse relationship types. The extracted sentence relationship dataset is named DSBATS-sn [1].

[1] https://drive.google.com/drive/folders/DSBATS

Contribution

- A new evaluation task for sentence embedding based on sentence relationships: Sentence Relationships Similarity Distinguishing (SRSD).

Relationship source: Word analogy

$$king : queen :: man : woman$$
$$dog : bark :: cat : mew$$
$$beach : sand :: ocean : water$$

Word analogy dataset

Bigger Analogy Test Set (BATS) (Gladkova et al., 2016)
A word analogy dataset organized as analogical clusters

- 20 categories of semantic relationships.
- Each category has 50 analogy pairs.

Word analogy dataset

Animal	Sounds
bee	buzz/hum
dog	bark/growl/howl/yelp/whine/arf/woof
cat	meow/meu/purr/caterwaul
duck	quack

Table: Excerpt from BATS datasets for the category E07 [Animal-Sounds] (Gladkova et al., 2016)

From word to sentence

A word analogy example from BATS (Gladkova et al., 2016).

beach : *sand* :: *ocean* : *water*

From word to sentence

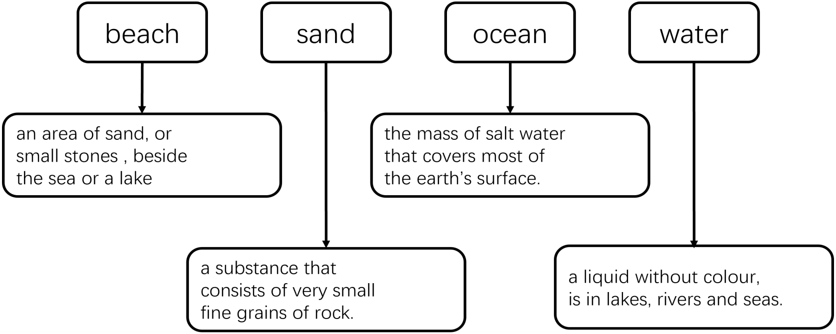


Figure: Word analogy relation from BATS and corresponding definitions from BabelNet (Navigli and Ponzetto, 2010).

From word to sentence

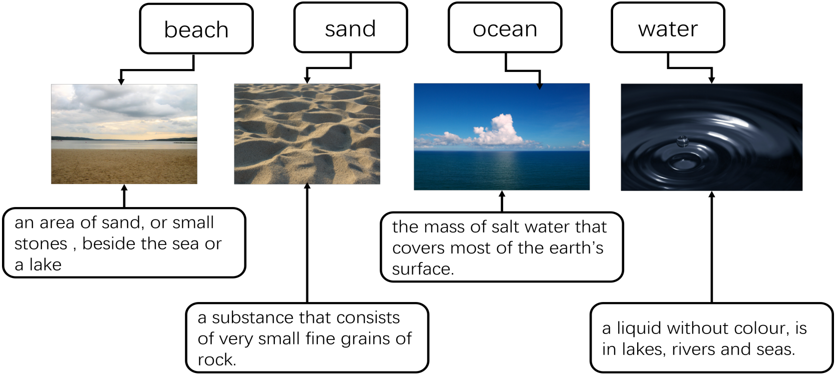


Figure: Words and definition sentences refering to the same concept. Pictures from Wikipedia.

Sentence source: Semantic Network

Language resource in network (graph) structure:

- Synset $\rightarrow$ Node
- Relation $\rightarrow$ Edge

Sentence source: Semantic Network

BabelNet (Navigli and Ponzetto, 2010): largest multilingual semantic network

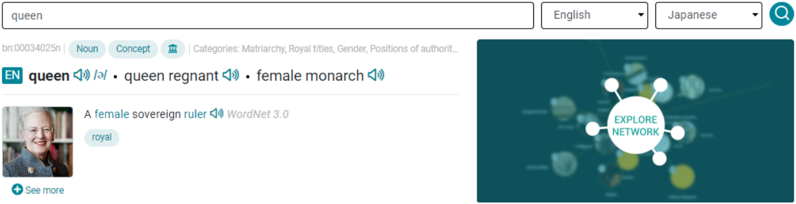


Figure: A synset from web version BabelNet (1)

Sentence source: Semantic Network

Figure: Web version BabelNet

Generation process

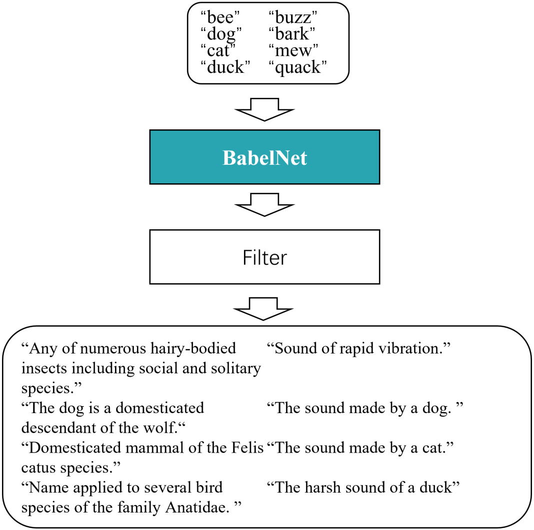


Figure: Input:word analogical cluster. Output: sentence pair cluster.

Generation process

Synsets: a set of synsets. One synset points to one concept, as well as a definition.

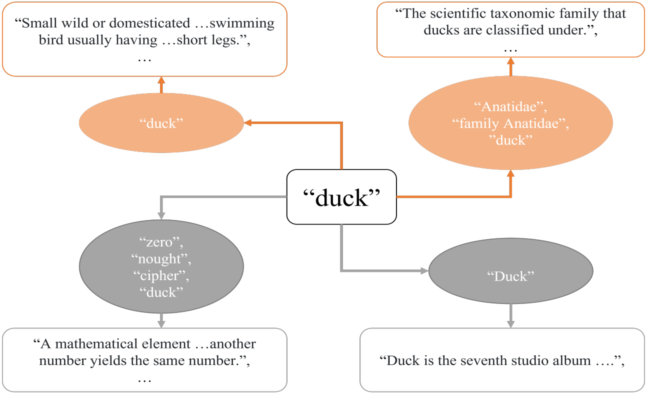


Figure: Synsets of "duck", orange synsets are kept.

Filtering process

- Deleting synsets with named entity
  Avoid the names of band, company, song, etc. In *king* : *queen* :: *man* : *woman*, Queen is not the famous band's name.
- Deleting synsets with capitalized words
  Avoid proper nouns. In *acrobat* : *troupe* :: *bird* : *flock*, Acrobat is not the name of a software from Adobe.
- Deleting synsets with lower synset degree.
  Avoid rarely used concepts.

DSBATS for Contrastive Learning: DSBATS4CL

DSBATS-sn: Definition Sentences from BATS with semantic network

“Any of numerous hairy-bodied insects including social and solitary species.”	“Sound of rapid vibration.”
“The dog is a domesticated descendant of the wolf.“	“The sound made by a dog. ”
“Domesticated mammal of the Felis catus species.”	“The sound made by a cat.”
“Name applied to several bird species of the family Anatidae. ”	“The harsh sound of a duck”

“an area of sand, or an area of sand, or small stones , beside the sea or a lake”	“small stones , beside a substance that consists of very small fine grains of rock.”
“the mass of salt water that covers most of the earth’s surface. “	“a liquid without colour, is in lakes, rivers and seas.”

Figure: 2 clusters extracted from DSBATS-sn

Contrastive learning

Purpose of optimization contrastive learning framework:
similar $\rightarrow$ close
dissimilar $\rightarrow$ far

Traditional contrastive learning loss

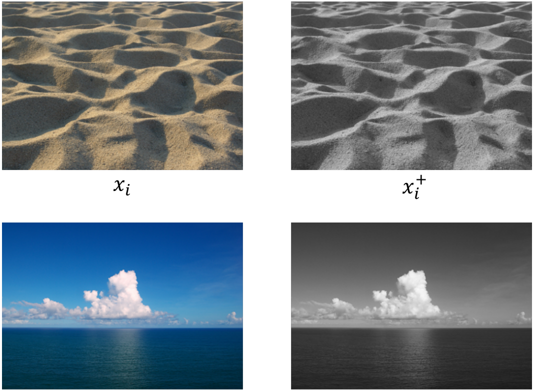
The loss function in contrastive learning is generally InfoNCE (van den Oord et al., 2018). In a batch of size S, the loss of the ith example is:

$$\text{loss}_i = -\log\left(\frac{e^{\text{sim}(x_i, x_i^{+})/\tau}}{\sum_{j=1}^{S} e^{\text{sim}(x_i, x_j^{+})/\tau}}\right)$$

similarity between positive examples

similarity between positive and negative examples

Traditional contrastive learning loss



x_i x_i^+

x_{i+1} $x^+_{(i+1)}$

Figure: Positive examples and negative sample in traditional contrastive learning in computer vision area

Data augmentation for DSBATS-sn

“Any of numerous hairy-bodied insects including social and solitary species.”	“Sound of rapid vibration.”
“The dog is a domesticated descendant of the wolf.“	“The sound made by a dog. ”
“Domesticated mammal of the Felis catus species.”	“The sound made by a cat.”
“Name applied to several bird species of the family Anatidae. ”	“The harsh sound of a duck”

“an area of sand, or an area of sand, or small stones , beside the sea or a lake”	“small stones , beside a substance that consists of very small fine grains of rock.”
“the mass of salt water that covers most of the earth’s surface. “	“a liquid without colour, is in lakes, rivers and seas.”

Figure: 2 clusters extracted from DSBATS-sn

Data augmentation for DSBATS-sn

"Any of numerous hairy-bodied insects including social and solitary species."

"Sound of rapid vibration." $\Big\}\ P_i$

"The dog is a domesticated descendant of the wolf."

"The sound made by a dog. " $\Big\}\ P_i^+$

"The sound made by a dog. "

"The dog is a domesticated descendant of the wolf." $\Big\}\ P_i^-$

Figure: An example of data augmentation

Problems with traditional InfoNCE with DSBATS-sn

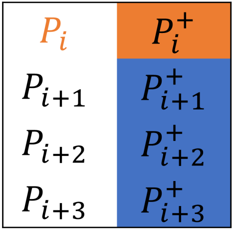


Figure: Positive and negative samples for p_i in InfoNCE for contrastive learning.

Problems with traditional InfoNCE with DSBATS-sn

"Any of numerous hairy-bodied insects including social and solitary species."

"Sound of rapid vibration."

P_i

"The dog is a domesticated descendant of the wolf.“

"The sound made by a dog. "

P_i^+

"The sound made by a dog. "

"The dog is a domesticated descendant of the wolf.“

P_i^-

Figure: An example of intra-cluster data augmentation

Problems with traditional InfoNCE with DSBATS-sn

“Name applied to several bird species of the family Anatidae. ”

“The harsh sound of a duck”

P_{i+1}

“Domesticated mammal of the Felis catus species.”

“The sound made by a cat.”

P_{i+1}^{+}

“The sound made by a cat.”

“Domesticated mammal of the Felis catus species.”

P_{i+1}^{-}

Figure: An example of intra-cluster data augmentation

Our loss

We only use the example as ${p_i}^-$ as negative example, different from InfoNCE. In a batch of size S, the loss of the ith example is:

$$\text{loss}_i = -\log\left(\frac{e^{\text{sim}(p_i, {p_i}^+)/\tau}}{\sum_{j=1}^{S} e^{\text{sim}(p_i, {p_j}^-)/\tau}}\right)$$

Our loss

P_i	P_i^+	P_i^-
P_{i+1}	P_{i+1}^+	P_{i+1}^-
P_{i+2}	P_{i+2}^+	P_{i+2}^-
P_{i+3}	P_{i+3}^+	P_{i+3}^-

Figure: Positive and negative samples for p_i in our contrastive learning.

English DSBATS-sn

We generated DSBATS-sn dataset for English for 20 categories.

Encyclopedic	Size	Lexicographic	Size
E01 country - capital	447	L01 hypernyms - animals	4318
E02 country - language	669	L02 hypernyms - misc	5005
E03 UK city - county	426	L03 hyponyms - misc	6768
E04 name - nationality	570	L04 meronyms - substance	1312
E05 name - occupation	912	L05 meronyms -part	854
E06 animal - young	566	L06 meronyms - part	4036
E07 animal - sound	633	L07 synonyms - intensity	1645
E08 animal - shelter	877	L08 synonyms - exact	1307
E09 things - color	934	L09 antonyms - gradable	5560
E10 male - female	384	L10 antonyms - binary	1453

Table: Size of English DSBATS-sn dataset.

Fine-tuning

Baseline models: BERT (Devlin et al., 2019), RoBERTa (Zhuang et al., 2021), SBERT (Reimers and Gurevych, 2019)
Training set: DSBATS4CL

Training set	Size
DSBATS4CL	2,244,530

Table: Size of DSBATS4CL

Intrinsic Evaluation

Task: Sentence Relationship Similarity Distinguishing (SRSD)

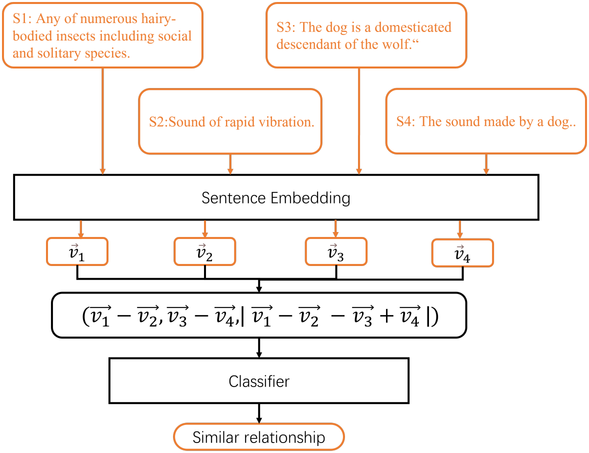


Figure: Input and output example of SRSD task

Intrinsic Evaluation

Test data: DSBATS-dic

Relationship source: BATS dataset

Sentence source: dictionary definitions from Oxford Dictionary, Merriam-Webster Dictionary, and Collins Dictionary.

Category	Size
L01	251
L02	225
L04	127

Table: The size of each category in DSBATS-dic

Extrinsic Evaluation

SentEval (Conneau and Kiela, 2018) is a tool that includes the following evaluation tasks for English.

Task	Description
STS	Semantic Textual Similarity, given a pair of sentences, calculate a similarity score for the two sentences
MRPC	Microsoft Research Paraphrase Corpus, given a pair of sentences, classify them as paraphrases or not paraphrases

Table: Introduction of extrinsic evaluation tasks

Evaluation results

		Intrinsic eval.	Extrinsic eval.	
Model	DSBATS4CL	SRSD	STS avg.	MRPC
BERT	w/o	58.18	18.63	68.81
	w/	**64.27**	**62.53**	**70.14**
RoBERTa	w/o	58.47	43.65	71.42
	w/	**65.83**	**65.11**	**71.83**
SBERT	w/o	61.68	62.84	73.51
	w/	**69.55**	**77.56**	**74.20**

After fine-tuning with DSBATS4CL, each model achieves better results on each task.

Conclusion

- Sentence relationships from word analogy contain world knowledge and improve sentence embedding quality. Result confirmed in English.
- The effectiveness of SRSD as an evaluation task: while the model works better on STS and MRPC, it also performs better on SRSD.

Future work

- Experiments on more low-resource languages
- Optimization of the filtering process for DSBATS-sn

Questions

Reference I

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations.
International Conference on Learning Representations (ICLR), 2018
Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and

Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, 2017 Anna

Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't.
In *Proceedings of the NAACL Student Research Workshop*, pages

8–15, 2016 Ruiqi Li, Xiang Zhao, and Marie-Francine Moens. A

Reference II

brief overview of universal sentence representation methods: A linguistic view.
ACM Computing Surveys (CSUR), 55(3):1–42, 2022 Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D

Manning. A large annotated corpus for learning natural language inference.
In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, 2015 Aäron van den

Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding.
CoRR, abs/1807.03748, 2018.
URL `http://arxiv.org/abs/1807.03748` Jacob Devlin,

Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT:

Reference III

Pre-training of deep bidirectional transformers for language understanding.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
doi: 10.18653/v1/N19-1423.
URL https://aclanthology.org/N19-1423 Alexis Conneau

and Douwe Kiela. SentEval: An evaluation toolkit for universal sentence representations.
In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki,

Reference IV

Japan, May 2018. European Language Resources Association (ELRA).
URL https://aclanthology.org/L18-1269 Roberto Navigli

and Simone Paolo Ponzetto. Babelnet: Building a very large multilingual semantic network.
In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225, 2010 Stefan Dumitrescu,

Andrei-Marius Avram, and Sampo Pyysalo. The birth of Romanian BERT.
In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online, November 2020. Association for Computational Linguistics.
doi: 10.18653/v1/2020.findings-emnlp.387.

Reference V

URL https://aclanthology.org/2020.findings-emnlp.387

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. *Chinese National Conference on Computational Linguistics*, pages 1218–1227, August 2021.
URL https://aclanthology.org/2021.ccl-1.108 Nils

Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks.
Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, 2019 Stefan Daniel Dumitrescu, Petru Rebeja, Beata Lorincz, Mihaela Gaman, Andrei Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia,

Reference VI

Cristina Iacobescu, et al. Liro: Benchmark and leaderboard for romanian language tasks.
In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021

Examples from DSBATS I

Word 1	Sentence 1	Word 2	Sentence 2
tomato	The tomato is the edible berry of the plant Solanum lycopersicum, commonly known as the tomato plant.	red	Red color or pigment; the chromatic color resembling the hue of blood
potato	Annual native to South America having underground stolons bearing edible starchy tubers; widely cultivated as a garden vegetable; vines are poisonous.	brown	Brown can be considered a composite color but is mainly a darker shade of red.
grass	A very large and widespread family of Monocotyledoneae, with more than 10.000 species, most of which are herbaceous, but a few are woody. The stems are jointed, the long, narrow leaves originating at the nodes. The flowers are inconspicuous, with a much reduced perianth, and are wind-pollinated or cleistogamous.	green	A colour sometimes referred to as Luggage or Luggage Green

Examples from DSBATS II

Word 1	Sentence 1	Word 2	Sentence 2
boy	A youthful male person.	girl	A female human offspring
brother	Son of the same parents as another person.	sister	Member of a non-Christian religious community of women.
bull	Intact adult male.	cow	Domesticated bovine animals as a group regardless of sex or age.